Vocal tract sagittal slices estimation from MRI midsagittal slices during speech production of CV

Ioannis K. Douros[1,2], Yu Xie[3], Chrysanthi Dourou[4], Jacques Felblinger[5],
Karyna Isaieva[2], Pierre-André Vuissoz[2], Yves Laprie[1]

[1] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2] Université de Lorraine, INSERM U1254, IADI, F-54000 Nancy, France
[3] Department of Neurology, Zhongnan Hospital of Wuhan University, Wuhan, 430071, China
[4] School of ECE, National Technical University of Athens, Athens 15773, Greece
[5] Université de Lorraine, INSERM 1433, CIC-IT, CHRU de Nancy, Nancy, F-54000, France

1 Synopsis

A transformation that connects the midsagittal dynamic slices of the vocal tract with the neighbouring frames was acquired at the image level. The transformation then was adapted to the midsagittal frames of a new speaker to give the synthesized neighbouring frames. The synthesized frames were compared with the original ones using image cross-correlation.

2 Purpose

Even though midsagittal slices are informative, there are many cases where they do not give enough information about what is happening to the non-midsagittal planes which are important from an acoustic point of view like in the case of /u/ or /i/. In this work we try to synthesize the images in the non-sagittal planes in order to have a better overview of the behavior of the articulators during speech production.

3 Introduction

Magnetic resonance imaging (MRI) is a modality widely used for speech studies because it has several advantages over other approaches like X-ray, EMA or ultrasound since it is non-invasive, non ionising method that can provide rich information about the vocal tract and its dynamics. With the advances in real time MRI (rtMRI) one can acquire 2D dynamic images with a good compromise between contrast spatial resolution and temporal resolution to study speech dynamics and analyze fast speech movements. Even though there are sequences to acquire 3D rtMR images [1] more progress is still required for the 3D rtMRI sequences to reach the quality level of 2D ones. However, information about the articulators outside the midsagittal plane is still required in order to better understand speech production. One way, is to use the rtMRI of the midsagittal plane and 3D volumetric information of static positions in order to estimate the 3D dynamics [2]. Such an approach provides interesting results, but this is quite hard to have a qualitative validation method that gives an idea of how good the estimation is. In this work we addressed this problem by acquiring 2D rtMRI data on the midsagittal and its neighbouring planes in order to directly synthesize non midsagittal frames for every time point and additionally develop a qualitative validation of the results using image cross-correlation between the original and the synthesized images.

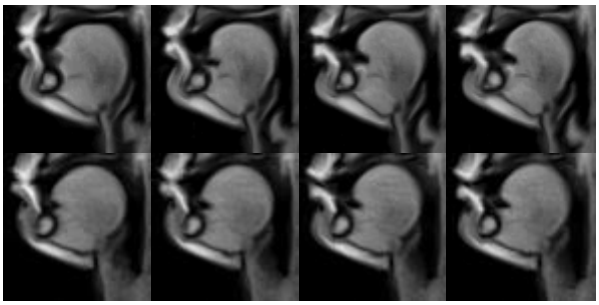
4 Materials and Methods

We chose a non-favourable case where the speakers where of different gender. One male and one female subjects were asked to repeat /pi/,/pa/,/pu/,/si/,/sa/,/su/ three times. Before the pronunciation of every CV, subjects were instructed to breath from the nose with closed mouth. Three sagittal planes (8mm slice thickness) were chosen for the acquisition, the midsagittal one and its left and right ones (with no space in between them). One plane was acquired per CV repetition. The data were acquired on a Siemens Prisma 3T scanner (Siemens, Erlangen, Germany) in CHRU of Nancy under the approved ethics protocol (ClinicalTrials.gov NCT02887053). The sequence used is real-time (50 fps) MRI Flash sequence [2]. In total 2000 images (including silence) were acquired. As a data preprocessing step, images should be labeled with their corresponding phoneme. Even though we could have used standard forced alignment phonetic labelling tools we did it manually to achieve a better temporal precision. The first step of the algorithm is to align the

data from the various planes of a given CV. For this purpose, piece-wise linear alignment was applied, using the midsagittal plane as a reference. A non-rigid image transformation $T_{t,M\text{-}L}$, $T_{t,M\text{-}R}$ of the train speaker that transforms the corresponding midsagittal frame to its left and right sagittal frame was calculated for every time frame. To compute the non-rigid image transformation a MATLAB function based on the algorithm described in [3] was used. The next step is to apply piece-wise linear alignment of the midsagittal frames between the train and the test speaker, using the test speaker as reference. For every time frame, a non-rigid transformation $T_{t,train\text{-}test}$, that transforms the midsagittal time frame of the train speaker to the corresponding frame to the test speaker, is computed. The last step is to adapt $T_{t,M\text{-}L}$, $T_{t,M\text{-}R}$ to the test speaker by composing these transforms using $T_{t,train\text{-}test}$ transform and apply the new transformation to the midsagittal frames of the test speaker in order to synthesize his/her left and right slices. To validate the results, cross-correlation between the synthesized and the original images [4], normalized by the autocorrelation of the original images, was used.

5 Results

Below we can see chosen images of the right side during /pu/ in both synthesized and the corresponding original form. Synthesized images have average correlation of 93.54% (± 1.06%) with the original ones, using normalized image cross-correlation. By visually inspecting the images, we can see that a small difference appears sometimes at the front part of the hard palate and some small differences in the eccentricity of the tongue. Apart from this point, images look quite similar in terms of vocal tract shape with an exception is a few cases were a small artifact may appear mainly at the front region of the tongue due to the existence of a similar artifact to the corresponding training images.



Selected right frames of /pu/. Top: synthesized images and bottom: corresponding original ones.

6 Discussion

One point is that since training and test speakers were of different gender thus with bigger anatomical differences than same gender subjects it is expected one to notice minor differences in the original and the synthesized images. A second point is that one can use standard automatic procedures to label the images at the preprocessing step using simultaneous audio recordings in which case our algorithm will run fully automatically. Further research could include exploring other ways to create the model or methods to better adapt it to a new speaker.

7 References

[1] Lim, Yongwan, et al. "3D dynamic MRI of the vocal tract during natural speech." *Magnetic resonance in medicine* 81.3 (2019): 1511-1520.

[2] Douros, Ioannis K., et al. "Towards a method of dynamic vocal tract shapes generation by combining static 3d and dynamic 2d mri speech data," 2019.

[3] Vercauteren, Tom, et al. "Diffeomorphic demons: Efficient non-parametric image registration." *NeuroImage* 45.1 (2009): S61-S72.

[4] Woo, Jonghye, et al. "A spatio-temporal atlas and statistical model of the tongue during speech from cine-MRI." *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6.5 (2018): 520-531.